EE 214 – Probability and Stochastic Processes

Machine Problem Set No. 2

(1st Semester 2019-2020)

1. The daily noon temperature in New Jersey in July can be modeled as a Gaussian random vector $\mathbf{T} = [T_1, \cdots, T_{31}]'$ where T_i is the temperature on the i^{th} day of the month. Suppose that $E[T_i] = 80°\text{F}$ for all i, and that T_i and T_j have covariance

$$\text{Cov}[T_i, T_j] = \frac{36}{1 + |i - j|}.$$

Define the daily average temperature as

$$Y = \frac{T_1 + T_2 + \cdots + T_{31}}{31}.$$

a) Based on this model, write a Python function julytemps(T), that calculates $P[Y \geq T]$, the probability that the daily average temperature is at least T degrees.

b) For the vector of daily temperatures $[T_1 \cdots T_{31}]'$ and the average temperature Y, estimate the probability of the event,

$$A = \{Y \leq 82, \min T_i \geq 72\}.$$

To form an estimate of (a), generate 10,000 independent samples of the vector $\mathbf{T}$ and calculate the relative frequency of A in those trials.

2. The disk of a computer system can be in three possible states: 0(Idle), 1(Read), or 2(Write). The values of the transition probabilities depend on factors such as the no. of sectors in a Read or Write operation and length of Idle periods. That is, consider the ff. state transition probabilities: Idle – Idle, $P = 0.95$; Idle – Read, $P = 0.04$; Idle – Write, $P = 0.01$; Read – Read, $P = 0.9$; Read – Idle or Write, $P = 0.05$; Note: Write has the same probability values as Read.

(a) Construct the Markov state diagram.

(b) Write a Python script markovdisk(n) to calculate the $n-$step transition probability matrix. Calculate the n-step transition matrices: $\mathbf{P}(10)$, $\mathbf{P}(100)$, $\mathbf{P}(1000)$.

3. In the attached IEEE ComSoc paper "Stock Market Forecasting using Hidden Markov Model: A New Approach" by Hassan and Nath. Answer the ff.:

(a) What is the observation probability density function? What do you think is the justification for using a 3-dimensional Gaussian distribution as the observation probability density function (obviously this is the state conditional output probability matrix **B**).

(b) What is the novel idea behind (from the authors) using HMM to predict the stock price?

(c) How is the training dataset used to estimate the HMM parameters (**A, B,** $\boldsymbol{\pi}$)?

(d) Provide specific examples on how the training dataset is used to estimate the HMM parameters.

(e) Discuss the results of this HMM based method as applied to some airlines stocks. Discuss the strong points and weak points of this HMM based algorithm to predict stock market trends with respect to the airlines stocks and with respect to this specific stock market location in general. What do you think,…would the same HMM based method yield the same quality of results if applied to other stock markets in other countries?

Note: The section on HMM in Lecture 6 Part 2 will be useful in understanding the HMM parameters mentioned in this paper.

4. Please refer to the attached article "Statistical Methods for Machine Learning" by Jason Brownlee. Based on this article, briefly answer the ff.:

(a) Why is Statistics important to Machine Learning? Site at least 10 examples where statistical methods are used in applied machine learning project.

(b) Give sample plots of simple data visualizations: Line Plot, Bar Chart, Histogram Plot, Box and Whisker Plot, Scatter Plot. Provide your own plots using the examples in the article as reference.

(c) Based on this article, differentiate the Law of Large Numbers from the Central Limit Theorem. That is, from the perspective of applied Machine Learning.

(d) In the section on Hypothesis Testing, how are Covariance and Correlation of random variables used?

(e) In the section on Estimation Statistics, what is the significance of knowing the Tolerance Interval for a Gaussian distribution and provide an example of how it is obtained. Provide your own example using the example in the article as reference.